BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 1 2005, pages 71–79
doi:10.1093/bioinformatics/bth463

A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data

Min Zou[1] and Suzanne D. Conzen[1,2,*]

[1]Department of Medicine 5841 South Maryland Avenue and [2]Committee on Cancer Biology, University of Chicago, Chicago, IL 60637, USA

Received on June 4, 2004; revised and accepted on August 3, 2004
Advance Access publication August 12, 2004

ABSTRACT

Motivation: Signaling pathways are dynamic events that take place over a given period of time. In order to identify these pathways, expression data over time are required. Dynamic Bayesian network (DBN) is an important approach for predicting the gene regulatory networks from time course expression data. However, two fundamental problems greatly reduce the effectiveness of current DBN methods. The first problem is the relatively low accuracy of prediction, and the second is the excessive computational time.

Results: In this paper, we present a DBN-based approach with increased accuracy and reduced computational time compared with existing DBN methods. Unlike previous methods, our approach limits potential regulators to those genes with either earlier or simultaneous expression changes (up- or down-regulation) in relation to their target genes. This allows us to limit the number of potential regulators and consequently reduce the search space. Furthermore, we use the time difference between the initial change in the expression of a given regulator gene and its potential target gene to estimate the transcriptional time lag between these two genes. This method of time lag estimation increases the accuracy of predicting gene regulatory networks. Our approach is evaluated using time-series expression data measured during the yeast cell cycle. The results demonstrate that this approach can predict regulatory networks with significantly improved accuracy and reduced computational time compared with existing DBN approaches.

Availability: The programs described in this paper can be obtained from the corresponding author upon request.

Contact: sconzen@medicine.bsd.uchicago.edu

INTRODUCTION

Genome-wide DNA microarrays are powerful tools, providing a glimpse of the signals and interactions within regulatory pathways of the cell. They enable the simultaneous measurement of mRNA abundance of most, if not all, identified genes in a genome under different physiological conditions. Because signaling pathways are dynamic events that take place over time, single time point expression profiles may not allow us to identify temporal events. This problem can be approached by performing a DNA microarray experiment with a series of time points following a physiological event.

Dynamic Bayesian network (DBN) analysis (Murphy and Mian, 1999; Imoto *et al*., 2002; Kim *et al*., 2003; Perrin *et al*., 2003) is well-suited for handling time-series gene expression data. To our knowledge, Murphy and Mian (1999) are to be credited with first employing DBN for modeling time-series expression data. In the DBN analysis, regulator–target gene pairs are usually identified based on a statistical analysis of their expression relationships across different time slices. For example, time slices T_1 for the regulator and T_2 for the target gene, where T_1 precedes T_2. The time period between the time slices of the regulator and target $(T_2 - T_1)$ is considered as the transcriptional time lag. Specifically, it is the time that it takes for the regulator gene to express its protein product and the transcription of the target gene to be affected (directly or indirectly) by this regulator protein. Consequently, we are more likely to observe a significant statistical correlation between the expression of a regulator and its target if biologically relevant time slices are used.

There are two major problems with current DBN methods that greatly reduce their effectiveness. The first problem is the lack of a systematic way to determine a biologically relevant transcriptional time lag, which results in relatively low accuracy of predicting gene regulatory networks. The second problem is the excessive computational cost of these analyses, which limits the applicability of current DBN analyses to a large-scale microarray data. Therefore, this paper introduces a DBN-based analysis that can predict gene regulatory networks from time course expression data with significantly increased accuracy and reduced computational time. Our approach differs from existing DBN methods [typified by Murphy's Bayes Net Toolbox (BNT) at

*To whom correspondence should be addressed.

www.ai.mit.edu/~murphyk/Software/BNT/bnt.html] in two major ways. First, in BNT, all the genes in the dataset are considered as potential regulators of a given target gene. In contrast, our method focuses on employing the biological fact that most transcriptional regulators exhibit either an earlier or simultaneous change in the expression level when compared to their targets (Yu *et al*., 2003). This limits the potential regulators of each target gene and thus significantly reduces the computational time. Second, in order to perform a statistical analysis of gene expression relationships, BNT generates a data matrix containing the time course expression profiles of the potential regulators and a given target gene. In this data matrix, the time course expression levels of all the potential regulators are aligned perfectly with each other throughout the time course. However, the expression levels of the target genes are misaligned with those of the potential regulators by one time unit. For example, the expression levels of the potential regulators at time point 1 are aligned with the expression level of the target gene at time point 2, where the time lag is just one time unit. Therefore, BNT automatically assumes that the time unit in a time course microarray experiment is the transcriptional time lag for all potential regulator–target pairs. This estimation of the transcriptional time lag can be inaccurate and results in a relatively low accuracy of predicting gene relationships using BNT. In contrast, our method proposes to use the time difference between the initial gene expression change of a potential regulator and its target as a reasonable estimation of the transcriptional time lag between these two genes, which can vary from zero (roughly simultaneous expression changes of the regulator and its target) to several time units. Based on these improvements, we expect that our DBN approach will uncover gene–gene relationships with a significantly increased accuracy and reduced computational time compared with existing DBN methods. The final steps in both our method and BNT are to calculate the conditional probabilities of the target gene expression in relation to the expression of its potential regulators, and subsequently the 'log marginal likelihood score'. Potential regulator(s) with the highest log marginal likelihood score will be ultimately selected as the final set of regulators for the given target gene. A conceptual representation of our approach is presented in Figure 1, and a detailed description of our method is presented in the Methods section.

MATERIALS: DATA AND SOFTWARE

To evaluate our approach, we report the analysis of the yeast cell cycle time-series gene expression data from Chou *et al*. (1998). This dataset has a large number of time points ($n = 16$) with relatively small time intervals (10 min), thus making it ideal for testing our approach. In addition, the yeast cell cycle has many previously established gene regulatory relationships (Simon *et al*., 2001), allowing ready confirmation of the accuracy of our algorithm-derived gene–gene relationships. For example, the Chou *et al*. (1998) yeast dataset contains 116 known cell cycle genes that encode either transcription factors (TFs) or their established targets. These genes can be inputted into our algorithm and predicted relationships are then verified by the comparison with established relationships.

Since Murphy's BNT already provides the necessary functionalities for building Bayesian networks, we implemented our new DBN analysis within the framework of BNT. The details of our approach are described in the 'Methods' section. The supporting programs to initially determine up- or down-regulation of individual genes and the transcriptional time lags between potential regulators and their targets are written in Java. These programs can be obtained from the corresponding author upon request.

METHODS

In this section, we describe the details of our DBN approach using the analysis of a set of hypothetical expression data as an example. This example includes four hypothetical genes A–D and their expression data at six evenly spaced time points T_1–T_6.

Step 1: Selection of potential regulators for each gene

We first determined the time points of the initial changes in the expression (up- or down-regulation) of genes A–D based up on their time-series expression data. Although there is currently no gold standard for determining what this threshold is for up- or down-regulation, we decided to use $\geq$1.2-fold (up-regulation) and $\leq$0.70-fold (down-regulation) compared to baseline gene expression as the cutoffs. Although these are relatively modest cutoffs, we did not want to miss genes with small, but potentially important changes in gene expression. We then determined the time points of the initial up- or down-regulation of genes A–D, and assigned genes with earlier or simultaneous changes in expression as the potential regulators of those genes with a later change in expression. In this way, we were able to select a subset of potential regulator genes for any given target gene.

The results of this potential regulator pre-selection for genes A–D are shown in Figure 2. Based on the criteria above for determining up- or down-regulated expression, the initial up-regulation of genes A and B occurs at T_2, gene C is initially up-regulated at T_3, and gene D is initially up-regulated at T_4. We selected genes A–C as the potential regulators of gene D because the initial up-regulation of genes A–C precedes that of gene D. This is followed by similar selection of potential regulators for other genes. In Figure 2, we illustrate a case of up-regulated expression, but similar potential regulator selection applies to down-regulated genes as well.

Step 2: Estimation of biologically relevant transcriptional time lag

After potential regulator selection, we next performed an estimation of the transcriptional time lag between potential

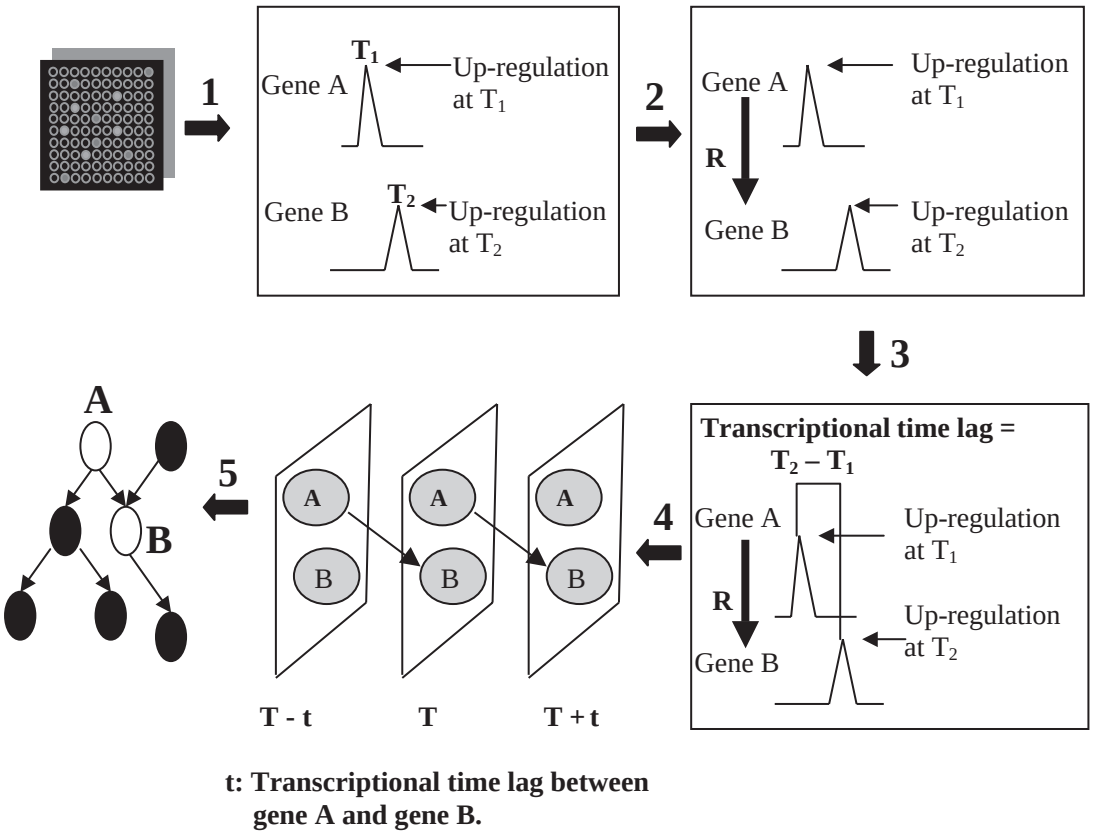

Fig. 1. A conceptual view of our DBN-based approach. **1**, Identification of the time point of the initial expression change (up- or down-regulation) of each gene based on the microarray time course expression data. **2**, Potential regulators are limited to those with simultaneous or antecedent expression changes when compared with their target genes (**R**, potential regulation). **3**, Estimation of the transcriptional time lag between the potential regulator and its target gene as the time difference between initial expression changes of these two genes. **4**, DBN: statistical analysis of the expression relationship between the potential regulator and its target gene in time slices that represent the transcriptional time lag between these two genes (as estimated in **3**). **5**, Predicted gene regulatory network.

regulators and their target genes. We propose that the time difference between the initial expression change of a potential regulator and its target gene represents a biologically relevant time period. This is expected to allow a more accurate estimation of the transcriptional time lag between potential regulators and their targets, because it takes into account variable expression relationships of different regulator–target pairs.

As an example, we illustrate the estimated time lags between target gene D and its potential regulators in Figure 3. We estimated that the time lags between potential regulators of genes A–C and their potential target gene D are two time units, two time units and one time unit, respectively. We then performed similar predictions for other potential regulator–target pairs. Of note, the transcriptional time lag estimated by our method can vary from zero to several time units. Since each target gene can have more than one regulator, we divided the potential regulators of each gene into different groups based on the individual transcriptional time lag with the target gene. As an example, we put genes A and B into one group since they have the same transcriptional time lag of two time units with respect to target gene D; gene C was placed in another group because it has a time lag of only one time unit with respect to gene D. The rationale for separating potential regulators into different groups is that different regulators may regulate the same target gene in either different time frames or in the same time frame. This allows us to analyze different potential regulators separately while grouping potential co-regulators together.

Step 3: Gene regulatory network modeling

The variables in our DBN analysis are the gene expression levels across different time points in the time course expression data. However, we did not use the absolute fold-change values, instead, we assigned '2' if the expression level is equal to or higher than the average expression level for that gene across all time points, and '1' if the expression level is lower than the average level. Note that we did not use the $\geq$1.2- or $\leq$0.70-fold cutoffs (see Step 1) to assign absolute up- or down-regulation to the expression level at each time point, instead we focused on the relative increase or decrease in expression levels. This is because the main focus of DBN is to identify the

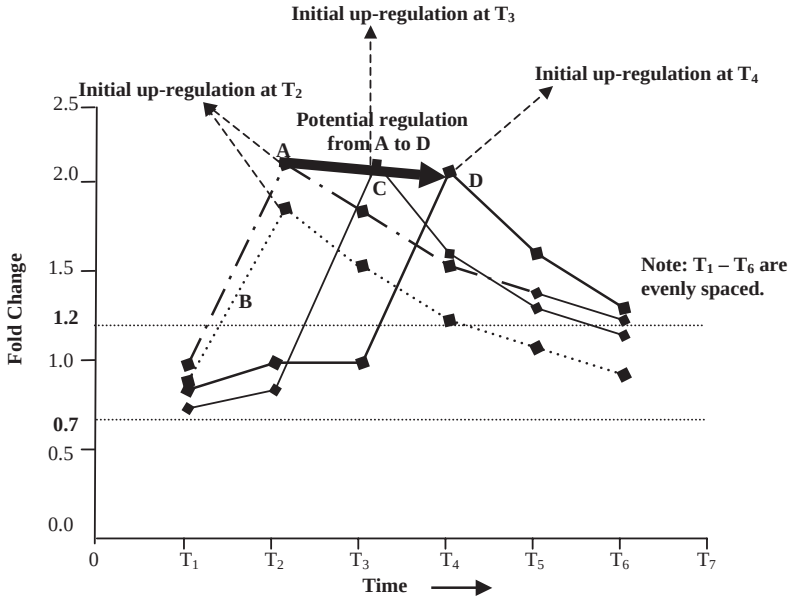

Fig. 2. Step 1: the dynamic expression profiles of genes A–D and the time points of their initial up-regulated expressions. Potential regulators are selected based on their simultaneous or antecedent expression change when compared with the expression change of their respective target gene.

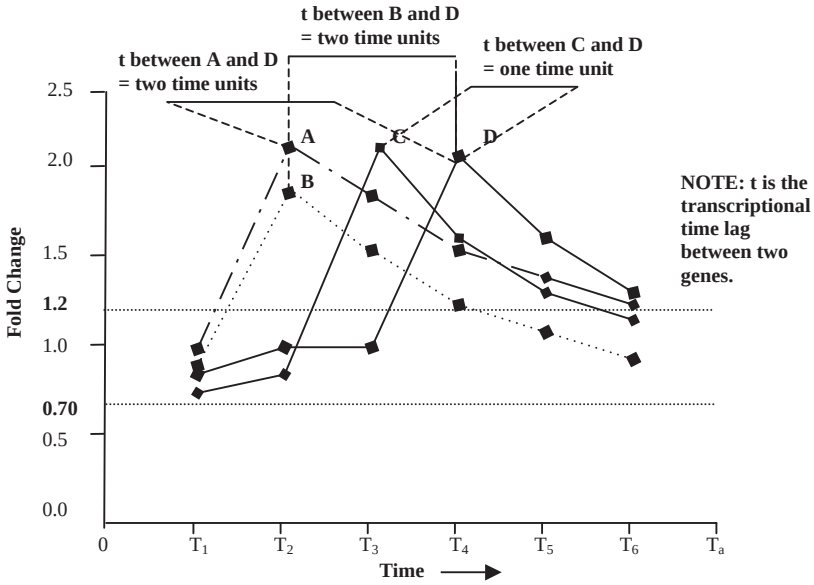

Fig. 3. Step 2: the transcriptional time lag between the potential regulator and its target gene is estimated as the time difference between the initial expression change of these two genes.

correlation between gene expression patterns, rather then their absolute expression value at any one particular time point.

We then used the results from Steps 1 and 2 to more accurately predict gene regulatory networks from time-series expression data, which are demonstrated by using the same example as in Steps 1 and 2. As stated in Step 2, we divided potential regulators of gene D into two groups based on their transcriptional time lags with gene D: a group of genes A and B with two time units as the transcriptional time lag with gene D, and gene C with one time unit as the time lag with gene D. For each group of potential regulators, we then generated all the subsets of this group, based on the user's pre-defined minimum and maximum number of regulators. This is because the number of co-regulators of a given target gene is unknown. The generation of the subsets of each group of potential regulators allows us to examine the expression

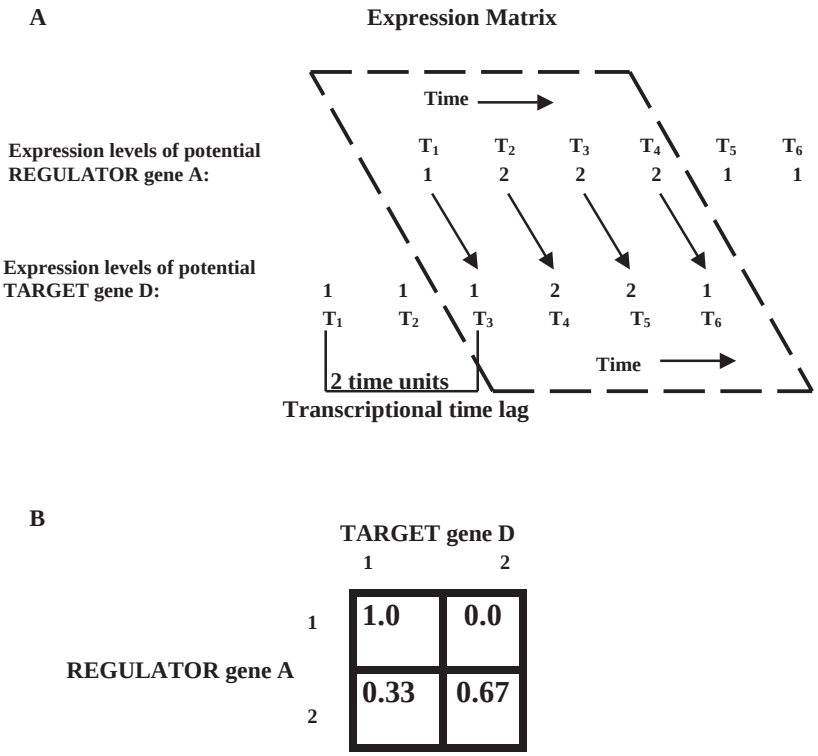

Fig. 4. Step 3: (**A**) Discrete expression values of potential regulator gene A and its target gene D are placed in a data matrix, where the expression level of gene A at time point T is aligned with the expression level of gene D at time point $T + t$ (t is the transcriptional time lag between genes A and D as estimated in Step 2). (**B**) The conditional probabilities of the expression of target gene D in relation to its potential regulator gene A are then calculated based on this data matrix.

relationships between all possible sets of co-regulators and their target gene. Therefore, the subsets of genes A and B are {gene A}, {gene B} and {gene A, gene B}; the subset of gene C is itself: {gene C}. Then, for each subset of potential regulators, we used the transcriptional time lag estimated in Step 2 to organize the expression data of the potential regulators and their target gene into an N × M matrix, where N is the number of the potential regulators plus the target gene, T is the number of time points in the original time series expression data, t is the estimated transcriptional time lag (represented by the number of time units) and M is the number of time points in the data matrix which is equal to $T - t$. Therefore, in this matrix, the expression value of the potential regulators at time T_1 are aligned with the expression value of the target gene at time $T_1 + t$, where t is the estimated transcriptional time lag. Note that the t-value (transcriptional time lag in time point units) may vary in different expression data matrices. In Figure 4A, we illustrate the data matrix for subset gene A with its target gene D.

After constructing expression data matrices of all target genes with their potential regulators, we calculated the conditional probabilities of each target gene in relation to its regulator genes based on the data matrices. The conditional probabilities of the expression of gene D in relation to the expression of gene A are shown in Figure 4B. Marginal likelihood scores were then calculated using these conditional probabilities. For each target gene, we then selected the subset of potential regulator(s) that gives the highest log marginal likelihood score as the final set of regulators for this target gene.

RESULTS

In DBN analyses of time-series expression data, at least two situations can exist. First, one might have some prior knowledge of the system studied, such as the identity of TFs in the system even though the targets of the TFs are unknown. Indeed, if we know the identity of the TFs in the system, we can use this prior knowledge to limit the potential regulators of each gene in the dataset to only these TFs, and then identify the targets of these TFs. In the second situation, we may not have any prior knowledge of the system, and thus need to identify regulatory networks by considering all potential regulator–target pairs. Therefore, in this work, we performed two sets of experiments to represent both of these possibilities.

In each experiment, we used both our approach and Murphy's BNT to analyze Chou *et al*.'s yeast cell cycle data, and compared the accuracy and the computational cost of the

Table 1. The results of Experiment 1 (incorporating prior knowledge of TFs)

Network	Correctly identified relationships	Misdirected relationship	Specificity (%)	Computational time (s)
(I) DBN_{our_1}	46	1	40	10
(II) DBN_{BNT_1}	18	4	11	60

Table 2. The results of Experiment 2 (no prior knowledge of TFs)

Network	Correctly identified relationships	Misdirected relationship	Specificity (%)	Computational time
(I) DBN_{our_2}	17	3	10	15 min
(II) DBN_{BNT_2}	8	7	3	8 h

two methods. In Experiment 1, we only allowed the nine TFs to be the possible regulators of the 116 genes in the dataset (including the nine TFs, because a given TF can be regulated by other TFs), and we identified the targets of these nine TFs. We denote the learned networks using our method and BNT as DBN_{our_1} and DBN_{BNT_1}, respectively. In Experiment 2, we excluded any prior knowledge of the yeast cell cycle, thus allowing all potential regulator–target pairs and subsequently identified the relationships between these 116 genes solely based on the time course microarray data. Regulatory networks identified using our method versus BNT in Experiment 2 are listed as DBN_{our_2} and DBN_{BNT_2}.

The results of Experiment 1 are summarized in Table 1, and the results of Experiment 2 are listed in Table 2. In both tables, row (I) represents the network identified by our method, and row (II) represents the network learned using BNT. 'Correctly identified relationships' specifies predicted relationships that have been established in yeast cell cycle regulation. 'Misdirected relationship' represents a gene relationship that is predicted to be in the reverse order of a known relationship. 'Specificity' is the percentage of correctly predicted known gene relationships out of the total number of predicted gene relationships. 'Computational time' is the running time of the analysis.

Experiment 1

Since we only allow the nine TFs to be the potential regulators in Experiment 1, the search space is relatively small and thus the computational time for both methods is relatively short. However, a close comparison of the computational times demonstrates a completion time of 10 s for our method and 60 s for Murphy's method (Table 1). The difference between the computational times is much more dramatic when the number of potential regulators increases as in Experiment 2.

In DBN_{our_1}, by selecting potential regulators based on their concurrent or antecedent change in expression in relation to the target genes, the number of misdirected relationships decreases from four (DBN_{BNT_1}) to one (DBN_{our_1}) (Table 1). Interestingly, the four misdirected relationships in DBN_{BNT_1} were all correctly reversed in DBN_{our_1}. Examination of the expression profiles of these four misdirected relationships reveals an earlier expression change of the known regulator gene compared with the target gene. For example, *SWI4*, a TF, is known to regulate gene *NDD1* (Simon *et al*., 2001). However, *SWI4* was erroneously determined to be the target of *NDD1* in DBN_{BNT_1}. Interestingly, using our method, *SWI4* becomes a regulator of *NDD1*. *SWI4*'s expression is up-regulated for 10 min compared with *NDD1*'s up-regulation for 20 min. However, *NDD1* apparently has a strong statistical relationship with *SWI4* based on their expression data, which results in the regulation of *SWI4* by *NDD1* in DBN_{BNT_1}. In our method, *NDD1* is excluded from being a potential regulator of *SWI4* because it has a delayed expression change compared with *SWI4*, and thus is not likely to be a regulator of *SWI4*. This results in assigning *SWI4* as a potential regulator of *NDD1*, and our statistical analysis confirmed that *SWI4* is a regulator of *NDD1*. Therefore, the misdirected relationship *NDD1* → *SWI4* in DBN_{BNT_1} was correctly reversed to *SWI4* → *NDD1* in DBN_{our_1}.

Interestingly, there is only one misdirected relationship in DBN_{our_1}. In this relationship, *SWI5*, although known to be a transcriptional target of *NDD1* (Simon *et al*., 2001), is predicted by our method to be a regulator of *NDD1* (*SWI5* → *NDD1*). This misdirected relationship is caused by the antecedent up-regulation of *SWI5* for 10 min compared with *NDD1*'s up-regulation for 20 min. Therefore, *SWI5* was selected by our method to be the regulator of *NDD1*, instead of vice versa. A further statistical analysis also confirmed the regulation directed from *SWI5* to *NDD1*. This finding indicates that even though most transcriptional regulators have either an earlier or simultaneous change in expression compared with their targets, there are exceptions. Earlier expression change of the target gene in comparison with that of the regulator gene may be caused by the different mRNA half-lives of the regulator gene and target gene, and may also suggest a feedback loop, where a target gene can in turn regulates its regulator. However, a further examination of Chou *et al*.'s yeast cell cycle time-series data showed that in 70% of the known yeast gene–gene relationships, the regulator gene has either an earlier or simultaneous expression change compared with its target gene. This suggests the general applicability of our method in discovering gene regulatory relationships.

The results from our analysis of yeast cell cycle expression data demonstrate that our method is capable of identifying a higher number of known gene–gene relationships compared with BNT. The number of correctly identified, already established gene–gene relationships increased significantly from 18 in DBN_{BNT_1} to 46 in DBN_{our_1} (Table 1).

Table 3. Correctly identified known gene–gene relationships in Experiment 1

Regulator	Target	Regulator	Target	Regulator	Target
(A) DBN_{our_1}					
SWI4	***PCL1***	SWI4	NDD1	FKH1	SWI6
SWI4	***CLN2***	FKH1	ACE2	SWI4	MBP1
SWI4	***OCH1***	NDD1	ACE2	SWI4	PCL2
SWI4	***HO***	SWI6	SIM1	SWI4	CLB6
MCM1	***STE6***	SWI4	FKS1	MCM1	SIM1
MCM1	***PIR3***	FKH2	GIC1	FKH2	CLB4
SWI4	***SWE1***	SWI4	SPT21	SWI6	CDC6
SWI4	***GIN4***	SWI4	RSR1	SWI6	AGA1
FKH1	***BUD8***	SWI4	CWP1	MCM1	MFA1
ACE2	***SPO12***	MCM1	CLN3	SWI5	MFA2
MCM1	CLN2	FKH1	UTR2	SWI4	CLB2
SWI4	RNR1	MCM1	GIN4	SWI4	PLB3
FKH1	HHF_1	SWI4	YBR071W	SWI5	YLR463
SWI6	HTB2	SWI6	YPR075	SWI6	SPO12
SWI5	EGT2	NDD1	CDC20		
SWI4	MNN1	SWI4	BUD4		
(B) DBN_{BNT_1}					
SWI4	***PCL1***	***ACE2***	***SPO12***	MBP1	CLB6
SWI4	***CLN2***	***SWI4***	***GIN4***	SWI4	HTA2
SWI4	***HO***	***FKH1***	***BUD8***	SWI6	RSR1
SWI4	***OCH1***	***MCM1***	***PIR3***	FKH2	ACE2
MCM1	***STE6***	SWI4	BUD9	FKH1	SWI5
SWI4	***SWE1***	SWI6	CIS3	SWI5	HSP150

(A) By our method. (B) By BNT. Relationships in italicized bold face type were identified by both our method and BNT. Relationships in normal font were identified by the corresponding method and not be the other method.

A close examination of the 36 gene–gene relationships correctly identified by our method (Table 3) and not by BNT reveals that all 36 relationships have a much stronger statistical correlation using the estimated transcriptional time lags compared with using the single time unit (10 min) as the time lag. As an example, *SWI4*, a TF, is known to transcriptionally regulate *MBP1* (Simon *et al*., 2001). However, this relationship was rejected by BNT because there is no significant statistical correlation between *SWI4* and *MBP1* expression when using a single time unit as the transcriptional time lag for the statistical analysis. This is illustrated by the low conditional probability of *MBP1*'s expression correlating with *SWI4*'s expression using the 10 minute time lag (Fig. 5A). However, when using a 30 min time difference (three time units) between the initial expression change of *SWI4* and *MBP1*, a strong statistical correlation was uncovered (Fig. 5B).

In contrast to the 36 known gene–gene relationships identified by our method exclusively, there are only eight relationships that are identified exclusively by BNT (Table 3). Three of these eight relationships have a better correlation if using 10 min as the transcriptional time lag compared with using the zero minute time lag estimated by our analysis. The other five relationships resulted from earlier expression changes of the target gene compared with their regulators.

Experiment 2

In this experiment, we compared the accuracy and efficiency of our method with BNT when no prior knowledge of yeast cell cycle TFs was inputted into the DBN model. This experiment allowed all the genes being analyzed to be potential regulators rather than only the nine TFs as in Experiment 1.

The difference between the computational cost of our method and BNT is even more dramatic in this experiment than in Experiment 1. The completion time for running our analysis on the dataset of 116 genes was only 15 min, while it took 8 h using BNT (Table 2).

As in Experiment 1, the number of misdirected relationships drops significantly from seven in DBN_{BNT_2} to three in DBN_{our_2} (Table 2). One misdirected relationship occurs in both networks. A close examination of the other six misdirected relationships in DBN_{BNT_2} reveals that these relationships include known regulators, which have an earlier initial expression change than their targets. Since the statistical analysis is the only measure to determine regulator–target pairs in BNT, the apparent statistical correlation erroneously determined the regulation from the known target gene to the known regulator gene in these six misdirected relationships. Interestingly, five of these six misdirected relationships were successfully reversed in our analysis. The other misdirected relationship was not reversed in our analysis due to an insignificant statistical correlation between the genes in this known relationship. Compared with the seven misdirected relationships in DBN_{BNT_2}, there are only three in DBN_{our_2}. The three misdirected relationships were caused by the fact that the known regulator has a later change in expression than its known target.

The advantage of using a more biologically relevant estimation of the transcriptional time lag is clearly reflected in these results. Twelve established relationships were correctly identified by our approach and not by BNT. Out of these 12 relationships, 10 have estimated transcriptional time lags other than 10 min, which is the arbitrary time lag used in BNT. In addition, all of the 10 relationships have much stronger statistical correlations when using their transcriptional time lags estimated in our method compared with using the 10 min time lag. Compared with the 12 known relationships identified in DBN_{our_2} (and not in DBN_{BNT_2}), there are only three that are identified exclusively in DBN_{BNT_2}. Two of the three relationships were not identified by our method because the known regulators in these relationships have later changes in expression than their targets. The other relationship was not identified by our method because it has a better statistical correlation if using a 10 min time lag compared with using the 20 min time lag we estimated in our analysis.

From the results of both experiments, we can see that the number of correctly identified known relationships decreases

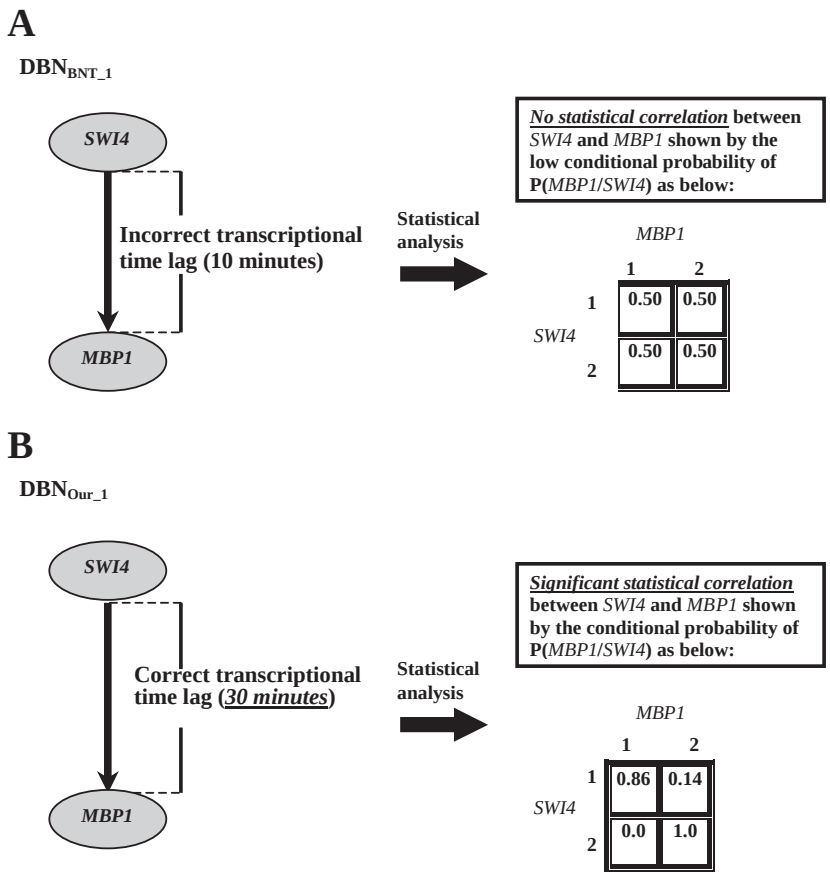

Fig. 5. (**A**) Known relationship *SWI4* → *MBP1* was not identified by BNT, due to the insignificant statistical correlation between these two genes as a result of using an incorrect transcriptional time lag (10 min = one time unit) for the statistical analysis. (**B**) Known relationship *SWI4* → *MBP1* was identified by our method, due to the significant statistical correlation between these two genes as a result of using a correct transcriptional time lag (30 min = three time units) for the statistical analysis.

in Experiment 2 when compared with Experiment 1. This finding reflects the fact that when we increase the number of potential regulators, the number of possible false positive predictions increases. However, if we reduce the number of potential regulators, we might miss uncovering interesting regulator–target pairs. This dilemma may be solved when we possess a more thorough understanding of the transcriptional regulation.

CONCLUSIONS

In this paper, we address two fundamental problems associated with current DBN analyses: (1) a low accuracy of predicted gene–gene relationships attributed to the arbitrary assignment of a transcriptional time lag and (2) an extremely long computational time due to the lack of an efficient approach to reduce the search space. In our approach, we consider the fact that gene regulators usually have either a simultaneous or antecedent changes in expression when compared to their targets. This consideration allows us to limit possible regulators of each gene thus reducing the search space. Furthermore, we use the time difference between the initial change in expression of a given regulator gene and its potential target gene to estimate the transcriptional time lag between these two genes. This estimation of transcriptional time lag increases the accuracy of predicting gene regulatory networks. In our current analysis, we used established TF–target relationships as measured by a genome-wide screen of promoter binding by tagged TFs to define correct gene–gene interactions (Simon *et al.*, 2001). Additional large-scale promoter binding screens have also been performed (Horak *et al.*, 2002) and are complementary approaches to a DBN-based analysis of global gene expression. Although assessment of the absolute predictive accuracy of any DBN method is limited by our current knowledge of established gene–gene interactions, our new method appears to be more accurate than traditional BNT in predicting gene–gene relationships.

In our analysis, we estimated the transcriptional time lag between a regulator gene and a target gene as the time difference between their initial expression change. The

time points of the initial expression change (up- or down-regulation) of a gene could be affected by the cutoffs we use to determine the significantly up- or down-regulated expression. Therefore, there can be error associated with the transcriptional time lag estimated by our method. Further work on how to most accurately determine thresholds of significant up- or down-regulation needs to be conducted.

Another important consideration is the noise in microarray gene expression data. One approach to deal with uncertainty in expression data is to perform replicate experiments of the same time course microarray experiment. Experiments can then be analyzed in at least two different ways. The first approach is to consider each dataset separately, which will result in independent gene regulatory networks for each dataset. We could then identify gene relationships that occur in the majority of the independently predicted networks. This approach may filter out false gene relationships that are caused by random noise, and therefore are not likely to occur consistently in different experiments. The second approach is to average the expression levels of each gene from independent replicates, and assign a standard error (SE) to each averaged value. We can then perform similar DBN analysis on this averaged data as described in the 'Methods' section. We could then use a scoring function that takes the standard errors into account instead of the 'log marginal likelihood' score that assumes the certainty of the expression values. We could redesign the scoring function so that gene relationships that have small standard errors for the averaged expression of its regulator genes and target genes will have a higher score, and will be given more weight than those gene relationships with higher standard errors. However, the first approach may be a better choice, because averaging may cause the loss of a significant relationship if the expression values of one experiment are particularly noisy.

Finally, we have shown that occasionally the expression of a target gene precedes that of the regulator gene. While this is an uncommon phenomenon, especially in eukaryotic cells, the occurrence of this phenomenon may be caused by several factors. The first factor is the variability of mRNA half-lives. For example, if the regulator gene's encoded mRNA has a significantly shorter half-life than that of its target gene, it may take the regulator's mRNA much longer to reach a significantly up- or down-regulated steady-state level when compared with the time it takes for the target's mRNA to undergo a significant change in steady-state expression level. This could lead to an apparently earlier time of threshold expression change of the target gene compared to its regulator gene. Taking variable mRNA half-lives into account is a current challenge for DBN developers. The second factor is the existence of gene regulatory feedback loops. The fact that a known target gene's change in expression occurs earlier than that of its known regulator gene may suggest a feedback loop, where the target gene can also regulate the regulator gene. However, a further examination of Chou *et al*.'s yeast cell cycle time-series data showed that in 70% of the known yeast gene–gene relationships, the regulator gene has either an earlier or simultaneous expression change compared with its target gene. This suggests a general applicability of our method in discovering gene regulatory relationships and providing testable hypotheses. Because many biological signaling networks involve key transcriptional events, this approach may be used to predict hypothetical gene regulatory networks from time course microarray data. For example, the transcriptional changes that follow growth factor or nuclear hormone receptor activation will lend themselves to this type of analysis.

ACKNOWLEDGEMENTS

We thank Dr Dan Nicolae and Dr William Hsu for their constructive suggestions regarding the development of our DBN approach. This work was supported by NIH grants CA90459, CA89208, ES0123282 and the Entertainment Industry Foundation.

REFERENCES

Chou,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. and Davis,R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Horak,C.E., Luscombe,N.M., Qian,J., Bertone,P., Spiccirrillo,S., Gerstein,M. and Snyder,M. (2002) Complex transcriptional circuitry at the G_1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.*, **16**, 3017–3033.

Imoto,S., Goto,T. and Miyano,S. (2002) Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. *Pac. Symp. Biocomput.*, **7**, 175–186.

Kim,S.Y., Imoto,S. and Miyano,S. (2003) Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief Bioinform.*, **4**, 228–235.

Murphy,K. and Mian,S. (1999) Modeling gene expression data using dynamic Bayesian networks. *Technical Report*, Computer Science Division, University of California, Berkeley, CA.

Perrin,B.E., Ralaivola,L., Mazurie,A., Bottani,S., Mallet,J. and D'Alche-Buc,F. (2003) Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **19**(Suppl.2), II138–II148.

Simon,I., Barnett,J., Hannett,N., Harbison,C.T., Rinaldi,N.J., Volkert,T.L., Wyrick,J.J., Zeitlinger,J., Gifford,D.K., Jaakkola,T.S. and Young,R.A. (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.

Yu,H., Luscombe,N.M., Qian,J. and Gerstein,M. (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.*, **19**, 422–427.